doi:10.11959/j.issn.1000-436x.2016055

支持技术创新的专利检索与分析

刘斌[1,2]，冯岭[1]，王飞[1]，彭智勇[1,2]
（1. 武汉大学计算机学院，湖北 武汉 430072；2. 武汉大学软件工程国家重点实验室，湖北 武汉 430072）

摘 要：介绍了目前专利检索和分析的主要研究工作，包括专利的可检索性、技术现状检索和相关性检索方法等，以及专利地图分析、新颖度分析和 PatentDom 专利分析框架等分析方法。最后基于深度学习的思想，讨论了新一代的支持技术创新的专利检索方法、专利论文检索方法以及专利趋势分析方法。
关键词：专利；专利检索；专利分析；深度学习
中图分类号：TP391.1 **文献标识码**：A

Patent search and analysis supporting technology innovation

LIU Bin[1,2], FENG Ling[1], WANG Fei[1], PENG Zhi-yong[1,2]
(1. School of Computer, Wuhan University, Wuhan 430072, China;
2. State Key Laboratory of Software Engineering, Wuhan Uvinersity, Wuhan 430072, China)

Abstract: The main research work of patent search and analysis were summarizes. The patent search includes patentability search,prior art search,and query expansion. And the patent analysis includes patent map, novelty analysis, and a new analysis framework named PatentDom. Finally, based on the idea of deep learning, three new methods of patent search and analysis are put forward.
Key words: patent, patent search, patent analysis, deep learning

1 引言

近年来，科学技术日新月异，经济全球化趋势增强，产业结构调整步伐加快，国际竞争日趋激烈[1~3]。知识或智力资源（包括专著、专利、商标、科技论文、技术报告以及科学实验数据等）的占有、配置、生产和运用已成为经济发展的重要依托，技术知识的重要性日益凸现。以知识为基础的产业在国内经济所占的比重不断提高，知识产权已成为国家之间、企业之间竞争的焦点。

专利是最典型的知识产权，也是数量最大的、增长速度最快的技术信息来源。美国专利申请始于 1790 年，中国则开始于 1985 年。表 1 反映了美国和中国的专利申请量的增长速度[1~3]。

表 1 专利发展趋势

国家	第 100 万件	第 200 万件	第 300 万件	第 400 万件	第 500 万件
美国	100 年	50 年	25 年	12 年	5 年半
中国	15 年	7 年半	2 年半	1 年	1 年

截至 2014 年底中国有效发明专利拥有量共计 66.3 万件，全世界范围内的专利数量已经达到 7 300 万件。根据世界知识产权组织的统计，专利文献中包含了世界上 95%的研发成果。如果能有效地利用专利信息，不仅可以缩短 60%的研发时间，还能节省 40%的研发经费[4, 5]。

专利蕴含着巨大的价值，吸引许多研究者的注

收稿日期：2015-10-10；修回日期：2016-01-20
通信作者：彭智勇，peng@whu.edu.cn
基金项目：国家自然科学基金资助项目（No. 61232002）；湖北省科技支撑计划基金资助项目（No. 2015BAA127）；武汉创新团队计划基金资助项目（No. 2014070504020237）
Foundation Items: The National Natural Science Foundation of China(No. 61232002), The Science and Technology Support Program of Hubei Province (No. 2015BAA127), The Wuhan Innovation Team Project (No. 2014070504020237)

意。2002年开始，日本国立情报学研究所在其举办的NTCIR会议设立专门的专利检索专题讨论会，并发布了若干专利测试数据集(如表2所示)，其中，NTCIR-3数据集包含跨语言检索任务。NTCIR-4,5,6数据集包含技术现状检索，专利分类等任务[6~8]。

CLEF(cross language evaluation forum)是面向欧洲语言的信息检索开放评测平台，从2009年开始设立专门针对专利检索的主题研讨会CLEF-IP，同时提供大约130万个英文专利，供研究者下载测试。

此外，一些重要的国际会议如CIKM、SIGIR等都设置了相应的专利Workshop，供研究人员进行交流。

表2 NTCIR数据集

会议	专利类型	时间范围	数量	主题数
NTCIR-3	日文专利	1998-1999	697 262	31
	英文摘要	1995-1999	1 700 000	31
NTCIR-4	日文专利，英文摘要	1993-1997	1 700 000	103
NTCIR-5	日文专利，英文摘要	1993-2002	3 496 252	1 223
NTCIR-6	英文专利(USPTO)	1993-2002	1 315 470	3 221

专利研究目前可以分为3类：1) 专利检索；2) 对专利文本进行各种深入分析；3) 与专利相关的其他研究，如推荐合作者[9]、专利续费等[10]。

2 专利检索相关评价标准与检索方法

2.1 专利检索评价标准

专利检索作为信息检索的一个分支，可以采用准确率和召回率对算法进行比较。但是准确率和召回率互相影响，理想情况下两者都要高[11]。一般情况下准确率高时，召回率就低；而召回率高时，准确率就低。专利检索侧重于召回率，为了更好地反映算法的全局性能，Magdy等[12]经过分析，设计专利检索评价值(PRES, patent retrieval evaluation score)。

$$PRES = 1 - \frac{\frac{\sum r_i}{n} - \frac{n+1}{2}}{N_{\max}} \tag{1}$$

其中，r_i是第i个相关专利文档的排名，n是专利文献集合中相关专利的数目，$N_{\max}$是用户最大检索的专利数。算法的$PRES$值越高，则召回率越高，且相关的文档排名越靠前。

专利检索按照检索目的可分为：可专利性检索（patentability search）也叫新颖性检索（novelty search）、专利技术现状检索（prior art search）、相关性检索等。专利的检索和一般的科技文献检索相比，有其特殊性，主要体现在以下4个方面。

1) 撰写方式的特殊性。论文撰写时，作者一般采用大家熟悉的描述方式，这样可以让读者更容易理解作者所要表达的含义。但是专利撰写时，申请人为了扩大自己所申请专利的保护范围和提高专利授权的可能性，往往使用一些模糊的术语和表达，甚至创造新的术语。

2) 对于专利检索，召回率比查准率更重要，因为如果漏检一条重要的专利，会给企业带来重大的损失。

3) 专利数据格式复杂。专利包含了分类号、权利声明等丰富信息。其中，专利分类号用来对专利文献进行分类，充分利用专利分类号等其他信息，可以使检索结果更准确。

4) 检索条件长度不同。对于专利申请人和专利审查员，他们更希望提供全文检索的功能，因此专利检索文本包含几百个关键字。而目前现有的一些检索比如即席检索(ad hoc search)、Web检索和文献检索的检索文本长度相对比较短，例如Google搜索的最佳长度为155个英文字符。

2.2 可专利性检索

因为专利检索的文本长度很大，所以缩短检索文本是一个简单可行的方法[13~18]。最常用的方法就是对专利文本各个词的频率(TF, term frequency)进行统计，选择Top-k高频词来代替原有查询进行检索。信息检索已有研究结果表明采用高频词来进行检索并不能得到很好的检索效果，因此提出了IDF(inverse document frequency)指数，并利用TF-IDF来计算每一个词的权重。然而，专利撰写者往往为了规避已有的技术，会创造一些新词，它们的TF-IDF值很高[11]。所以采用TF-IDF方法仅能检索到少量的相关专利。Hideo等[13]针对跨库检索提出了一种词过滤的技术，每个词被赋予一个过滤权重TDV(term distillation value)。

$$TDV = QV \cdot TV \tag{2}$$

其中，QV表示词在查询条件中的重要性，TV表示词在目标语料库中的重要性。假设p为词在查询条件中出现的概率，q为查询词在目标集合中出现的

概率。对于一个词，概率 p 可以利用标题(t)和摘要(a)中词的频率进行计算，计算方法如式（3）所示。

$$p=\frac{n_t}{N_p} \text{ 或 } p=\frac{n_a}{N_p} \tag{3}$$

其中，n_t 是专利标题中包含该词的专利数量，n_a 代表专利摘要中包含该词的专利数量，N_p 代表集合中的专利总数量。

概率 q 利用目标集合(i)和整个专利文档的词(w)的分布进行计算，计算方法如下

$$q=\frac{n_i}{N_i} \text{ 或 } q=\frac{n_w}{N_p} \tag{4}$$

其中，n_i 是目标集合中包含该词的文章数，N_i 是目标集合总的文章数，n_w 是专利中包含该词的专利数量，N_p 代表 NTCIR-3 中专利数量。

对于一个检索词，论文依据不同的规则设计了 9 种计算 QV 的方法，以及 5 种计算 TV 的方法。采用 NTCIR-3 的数据作为测试集，该算法效果排名第 1（如表 3 所示），表明该方法可以有效地进行跨库检索。表 3 中 QV_2 的含义是一个词的 QV 值等于该词的频率。QV_0 含义是 QV=1，即 TDV 的值仅依赖于 TV。$TV_3=\frac{ap}{ap+(1-a-e)q+e}$，$a$ 和 e 是预先定义好的常量，P@10 为前 10 个专利准确率。

表 3　　NTCIR-3 数据集测试结果

QV	TV	p	q	MAP	P@10
QV_2	TV_3	t	i	0.279 4	0.390 3
QV_0	TV_3	t	i	0.270 1	0.348 4
QV_2	TV_3	a	i	0.268 8	0.364 5
QV_5	TV_3	a	w	0.263 7	0.361 3

审查员（或者发明人）通过输入待审核专利的权利声明（claim），算法自动抽取相关的关键词进行检索，返回相应的文档，进而判断权利范围要求的合法性。从表 3 中可以看出，算法的平均准确率（MAP）小于 0.3，在排名前 10 的专利准确率不超过 0.4。这是因为专利中存在大量语义含混不清的词，导致词过滤技术方法面临较大的挑战。

2.3　技术现状检索

技术现状检索就是给定一个技术背景（如一篇专利），找出与其相关的专利。技术现状检索可以帮助公司掌握最新相关领域的发展现状，辅助公司确定新的开发领域，合理分配宝贵的资源。检索条件的抽取是技术现状检索成功的关键，由于专利检索更注重召回率，采用查询扩展是比较有效的方法，所以寻找有效的扩展词就成为研究的重点。

2.3.1　第三方知识库的扩展方法

专利现状查询面临 2 个挑战：1）由于输入为一组关键词，而各个关键词可能属于不同的主题，因此无法表达一个准确的查询需求。2）查询中常常存在歧义词，如“苹果”可能表示苹果公司，也可能表示水果。信息检索已有研究表明，借助于维基百科这样的公共知识库可以提高检索的准确率和召回率。IPC 分类是国际通用的专利分类方法，它描述该类专利的特点、功能，因此可以把 IPC 分类描述看成是一种知识库，借助于 IPC 可以用来进行语义消歧，提高专利检索的准确率和召回率。例如，当“苹果”出现在电子分类的 IPC 下时，它通常指的是苹果公司，当出现在农业和林业等分类下时，苹果可以看成是水果[19, 20]。

Mahdab[17]利用 IPC 描述作为扩展词典，提出了一种基于位置近邻的查询扩展方法，并对检索结果进行重排序，从而提高检索的准确率和召回率，算法步骤如下。

1) 对于被检索的专利，使用第一条权利声明代替整个专利作为查询条件。

2) 提取 IPC 文本中专利特征的相关性描述，去除专利领域的停用词，建立候选扩展词表。

3) 对专利库中的每一条专利，计算扩展词和查询词的相关度，选择 Top-k 个相关度最高的词作为查询扩展词。扩展词和查询条件相关度计算方法如下

$$P(q\mid i,d)=\sum_{j=1}^{m}P(q\mid t_j)P(j\mid i,d) \tag{5}$$

其中，$P(q|t_j)$是查询词 t_j 在查询条件中出现的概率；j 是查询词在专利文档 d 中的位置。$P(q|t_j)$可以采用最常见的词频统计的方法计算。$P(j|i,d)$用来计算专利文档中第 i 个位置是扩展词与第 j 个位置是查询词的相关性概率。它的计算可以采用位置核函数来进行计算，如高斯距离核函数、拉普拉斯距离核函数等。该公式的含义是查询词在查询条件中出现的概率越大，扩展候选词离查询词在文中位置越近，它们的关系就越紧密，则该权重越大。

4) 利用查询扩展词进行查询，并对查询结果利用式(6)重新计算专利相关度。

$$P(q\mid d,e)=\sum_{i=1}^{|d|}P(q\mid i,d)P(i\mid d,e) \tag{6}$$

其中，$|d|$代表专利文档 d 中词的总数，$P(i|d,e)$表示第 i 个词是扩展词的概率。如果第 i 个词是扩展词，那么它的概率是扩展词所有出现位置总数的倒数，否则概率为零。

以 CLEF2010 作为实验数据，该方法效果如表 4 所示，和其他方法相比，检索的准确率有了较大的提高（8%）。主要原因是专利申请人在撰写专利时都要参考和使用 IPC 的描述信息，因此利用 IPC 作为扩展词可以最大限度地把扩展词的歧义降到最低；此外，计算相关度时将词的分布和位置结合起来。

表 4 IPC 扩展检索对比

算法	*MAP*	*PRES*	召回率
IPC 扩展(第一声明)	0.129 3	0.514 0	0.606 7
IPC 扩展(所有声明)	0.144 5	0.491 1	0.562 4
Ganguly[21]	0.127 8	0.460 4	—
Magdy[22]	0.139 9	0.486 0	—

2.3.2 基于主题的检索

专利作为一种文档，必然包含一定的主题。判断 2 个文档相似性的常规方法是通过统计 2 个文档中共同出现的单词数，这种方法没有考虑到文字背后的语义关联，可能存在 2 个文档共同出现的单词很少，但 2 个文档是相似的情况。LDA 模型可以提高检索的准确性，因此在信息检索和自然语言处理中得到了广泛的应用。

Krestel 等[23]将 LDA 模型应用到专利推荐，提出了基于潜在主题的专利推荐方法。根据专利的特点，将专利分成 5 个部分：题目（title）、摘要（abstract）、权利声明（claims）、概要（summary）和具体实施（details），利用 DMR（dirichlet multinomial regression）对专利和查询条件进行计算，选择相似度高的专利进行推荐，具体方法如下。

1) 对于一个给定的专利 q，利用 TF-IDF 从专利集合中选取 Top-k 个内容相关的专利，生成初始候选集。

2) 对于 Top-k 个专利，分析专利引用部分，如果该专利引用了其他专利，将这些被引专利加入到候选集合中。

3) 对候选集中的每一个专利 d，按照下面的方法计算 *DMR* 值。

$$DMR_d = p_{dmr}(q|d)\prod\nolimits_{w\in q} P_{dmr}(w|d) \tag{7}$$

$$P_{dmr}(w|s) = \sum\nolimits_{s=1}^{4} \frac{N_s}{N_d} P_{dmr}(w|s) \tag{8}$$

$$P_{dmr}(w|s) = \sum\nolimits_{z}^{N_z} P(w|z,\hat{f})\ P(z|\hat{q},s) \tag{9}$$

式(9)中 z 是专利包含的主题数，取值为专利总数的开方($\sqrt{N_c}$)。N_s 是每一个部分词的总数，N_d 是专利 d 包含的总词数。$\hat{f}$ 和 $\hat{q}$ 是词和主题的后验概率估计，可以通过 Gipps 抽样的方法进行计算。

该方法随机选择了 2012 年 12 月 3 日发布的 100 个专利，对每一个专利选择 500 个相似度最大的专利，加上被引专利得到一个包含 27 500 个专利的集合。表 5 是将该方法和 BM25、语言模型(LM)进行比较的结果。

表 5 基于 LDA 的专利检索对比

算法	*MAP*	
	平均准确率	提升（+）/下降（-）
BM25	0.062	- 51.2%
LM	0.127	—
LDA	0.134	+5.5%
DMR	0.143	+12.6%
LM-DMR	0.164	+29.1%
LM-LDA	0.177	+39.4%

LM 主要考虑词的分布，LM-DMR 和 LM-LDA 方法是用 DMR、LDA 对语言模型进行扩展。以 LM 为基准，可以发现利用主题可以提高平均准确率，将语言模型和主题模型进一步结合使检索结果更精确，这也符合一般的规律。

2.3.3 基于引用关系的查询扩展方法

专利申请书还包含了丰富的引用信息。Mahdabi[24]对专利文档进行实验观察，发现 2 个专利之间的相关性不仅与两者的文本相似度相关，而且很大程度上与两者之间的引用关系相关，进而提出基于时间感知的加权 PageRank 算法 AQE-TPR，具体步骤如下。

1) 查询专利集合，得到 Top-k 个文本相似度高的专利作为根集合，然后找出所有引用该 Top-k 个专利以及 Top-k 所引用的专利，根据引用关系构建专利引用网络。

2) 对其中每一个节点按照式（10）计算其初始概率。*age* 是专利授权时间，t_d 是时间间隔因子，专利授权越早和查询条件相关的可能性越低。

$$P_i = e^{\frac{-age}{t_d}} \quad (10)$$

如果专利 i 引用专利 j，那么 i 和 j 之间就存在一条边，边的权值对应于专利 i 和 j 之间的关联程度。AQE-TPR 综合考虑专利 i 和 j 的 IPC 分类、内容相似度，发布时间间隔、共同发明人、共同的专利权人。当组合权重大于 0.5 时，a_{ij}=1，反之 a_{ij}=0。这样就构成一个专利引用网络 G_{cit}，利用 PageRank 算法计算每一个专利的值。

3) 计算每一个词的权重，方法如下

$$P(t \mid G_{\text{cit}}) = \sum_{D \in G_{\text{cit}}} P(t \mid D) P(D) \quad (11)$$

其中，$P(D)$对应于专利 D 的 PageRank 值，$P(t|D)$是该词在专利 D 中出现的概率。如果一个专利的 PageRank 值越大，那么该专利处于核心地位，如果一个词在很多专利中出现，那么该词非常重要。

4) 综合考虑查询条件 Q_{orig} 和专利引用网络，利用式（12）计算扩展词的概率，选择 Top-k 个概率最高的词作为扩展词，$?$ 是预先定义的常数。

$$P(t \mid Q) = l\ P(t \mid Q_{\text{orig}}) + (1 - l\ ) P(t \mid Q_{\text{cit}}) \quad (12)$$

以 CLEF-IP2011 的数据为实验对象，比较结果如表 6 所示。可以发现 AQE-TPR 方法好于 Nijm 和 Hyder 算法，Nijm 和 Hyder 算法在 CLEF-IP 2011 比赛中排名第一和第二。本方法的贡献就是通过 PageRank 算法综合考虑专利各个部分的信息，从而提高检索的准确率和召回率。

表 6　　基于引用的专利检索对比

算法	*MAP*	召回率	*PRES*
Nijm(rank1)	0.058 2	0.630 3	—
Hyder(rank2)	0.059 3	0.571 3	—
AQE-TPR	0.125 0	0.647 0	0.536 3

此外还有一些方法利用查询扩展提高专利检索的召回率[25,26]。Bashi[27,28]利用词语位置计算语料库和查询条件的相关性，并选择最相关的若干文档，利用伪相关反馈进行查询扩展。Bhatia[26]将专利文献分割成大小相同的片段(snippet)，并将查询条件分割成较小的句子。将查询条件和专利文献进行比较，选择相似度最大的片段，并以此返回相应的专利文献，该方法能提高查询的响应时间。较早的方法有 Hironori[25]提出的利用聚类进行查询扩展，该方法将专利聚类成一个层次结构，在不同的层次上进行查询扩展以提高召回率。

2.4　相关性检索

专利相关性是指该专利和哪些专利相关。英文专利包含专利之间的引用关系。和论文的引用关系不同，专利对其他专利的引用意味着本专利的权利声明受到限制，即本专利的价值会变得更低，所以专利发明人在引用其他专利时会显得非常“小气”，这对已有的专利是非常不公的[29,30]。

Sooyoung 等[29]提出了基于价值驱动的专利引用推荐方法 CV-PCR。CV-PCR 将专利 D_i 表示为一个五元组$<T_i,C_i,V_i,a_i,R_i>$，其中，T_i 代表专利的文本内容，C_i 代表专利的 IPC 代码，V_i 是专利发明人，a_i 是专利所有权人，并以此构建专利异构信息网络，如图 1 所示，网络中边的含义如表 7 所示。

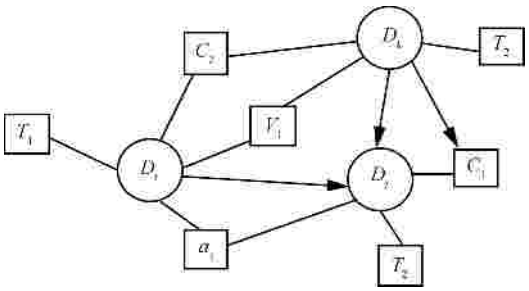

图 1　专利异构信息网络

表 7　　网络拓扑含义

边	含义
$<D_i,D_j>$	D_i 引用 D_j
$<D_i,C_2>$	C_2 是 D_i 的分类号
$<D_i,V_1>$	v_1 是专利 D_i 的发明人
$<D_i,a_1>$	a_1 是专利的所有权人
$<D_i,T_1>$	D_i 包含关键词 t_1

对于一个给定的专利 D_q，CV-PCR 分为 3 步推荐相关的引用。

1) 采用常规的专利检索方法检索出若干个相关的专利，并计算专利相关度。

2) 以专利异构信息网络为基础，根据式（13）计算专利的特征值。特征包括：专利之间是否有引用关系、专利的相似度、专利主分类号、专利次分类号、专利发明人、专利权人、专利内容。

$$P(D_b \mid x) = \frac{|\{D_b : x \in X(D_a), cite(D_a, D_b)\}|}{|\{D_a : x \in X(D_a)\}|} \quad (13)$$

其中，分母的含义是对于专利 D_b 以及它的特征 x，有多少专利具备特征 x；分子的含义是这些专利中同时引用专利 D_b 的数量。

给定一个查询专利 D_q，对网络中每一个专利计算所有特征值的平均值，方法如下。

$$VDF(D_{cand}, X(D_Q)) = \frac{\sum_{x \in X(D_Q)} P(D_{cand} \mid x)}{|X(D_Q)|} \quad (14)$$

3) 对这些专利采用基于监督排序学习算法(RankSVM)进行重新排序。

CV-PCR 和 BL1 方法和 BL2 方法进行了对比[19,31]。其中，BL1 方法是基于排序学习的相关专利检索，BL2 方法是科技论文引文推荐方法，比较结果如表 8 所示。通过比较可以发现该方法在推荐相关专利方面具有优势。主要原因是该方法不仅考虑了专利的内容，而且考虑了专利的其他有用信息。这进一步说明了专利检索有其特殊性，不能简单地照搬传统的信息检索方法。

表 8 专利相关性检索对比

方法	*MAP*	召回率		
		R@10	R@50	R@100
CV-PCR	0.357	0.214	0.453	0.524
BL2(CTM)	0.318	0.19	0.346	0.406
BL1	0.173	0.131	0.272	0.347

3 专利内容扩展分析方法

专利分析是对专利说明书或者专利公报中大量专利信息进行分析、加工、组合，并利用统计学的技巧和方法使这些信息转化为具有总揽全局及预测功能的竞争情报，从而为企业技术、产品及服务研发提供决策参考。常见的专利分析有：专利地图、专利价值计算、专利新颖性分析等。

3.1 专利地图

专利地图(patent map)是采用统计分析方法加以缜密及精细剖析整理制成的各种可分析解读的图表信息，具有类似地图指向功能。专利技术功效地图通常将专利分解成技术手段和技术效果 2 个维度，制作成矩阵或图表，横轴代表一项技术，而纵轴代表技术效果[32,33]。

图 2 是对手机领域从 2002 年~2007 年专利申请进行划分得到的专利技术功效矩阵，从中可看出，每一年手机功效的发展趋势，例如，2002 年多媒体技术、智能化技术和时尚外观设计催生了手机中的照相功能。图 2 中包含 3 个技术空白区。如技术空白区 2 表明手机产业中外观设计发明不多，还有很大的发展空间；空白区 3 表明多媒体、智能化和数据连接技术在手机 GPS 导航中运用还不多[34,35]。

功效	照相	黑莓机	Wi-Fi双模	手机电视	GPS导航	智能性
年份 / 技术	2002年	2003年	2004年	2005年	2006年	2007年
多媒体网络		√	√		√	
多媒体技术	√		√	√	空白区3	
智能化技术	√	空白区1				√
数据连接技术		√		√		
时尚外观设计	√	空白区2				

图 2 2002 年~2007 年手机专利地图

其实，从最近几年手机的发展趋势可以看出，外观设计已经成为手机一个很重要的卖点，且目前的手机都具备 GPS 导航功能，导航中各种语音提示，以及近乎真实三维地图、实时路况信息以及周边相关的娱乐、餐饮、住宿等信息都已经有效地集成到导航中。所以好的专利地图可以帮助用户快速了解领域技术现状、发现技术真空，对指导专利研发有着重要作用。目前专利地图的制作仍然采用半人工半自动化的过程。例如对于专利技术/功效矩阵图，因为技术和功效通常很难区分，所以提取一篇专利中技术与功效往往是一件非常难的事情。此外，专利的数量过于庞大，且所属的领域具有很大的差异[26]。

3.2 新颖性分析

专利新颖性并没有一个公认的定义。一般可以这样理解专利的新颖性，新颖性是指发明不属于现有技术，也没有任何单位或者个人就同样的发明向专利局提出过申请，并记载在申请日以后（含申请日)公布的专利申请文件或者公告中。

Hasan 等[36]提出了一个利用词新颖度计算专利新颖度的方法，并设计一个专利排序系统 COA (claim originality analysis)，针对专利的价值（包括专利的新颖性和重要程度），对专利进行排序。

COA 方法基于专利的总体贡献度对专利进行排序。总体贡献度是该专利所有关键短语的贡献度之和，总体贡献度越大，代表该专利所具有的价值越大，具体步骤如下。

1) 提取专利文本的关键词，COA 采用自然语言处理方法中的 n 元语法（n-gram）从专利文本中提取所有短语。在关键短语识别部分，作者构建了背景词典，将出现频率大于 k 的短语放入背景词典。

通常，这些短语出现频率较高，但对专利的价值贡献较小，所以将这部分短语过滤掉。经过以上 2 个部分，剩下的短语被识别为关键短语。同时，COA 引入了时间窗口的概念，仅考虑最近 k 年内新出现的短语，进一步减少了关键短语的数量。

2) 计算关键短语贡献度。在 COA 中，关键短语的贡献度基于 2 个方面：关键短语的频度和短语出现的时间长度。贡献度值的大小与关键短语出现的频度成正比，与短语出现的时间长度成反比。

3) 计算专利的价值。COA 设计了 2 种专利价值计算方法：①对每条专利的所有关键词的贡献度进行线性累加，得到该条专利的总体贡献度；②将关键短语的数量作为专利的价值。

该方法以 IBM 申请的专利为实验数据进行效果评估。首先采用领域专家对每一个申请的专利人工分为 3 类：1 核心(excellent)，2 重要(good)和 3 一般(not-so-good)。作者然后采用 COA 方法对专利进行打分，并和人工分类的结果进行比较，比较结果如表 9 所示。从表中可以看出属于类 1 专利的 COA 值远远大于属于类 3 专利的值。

表 9　　COA 对专利打分结果

类别	数据集 1		数据集 2		数据集 3	
	COA(1)	COA(2)	COA(1)	COA(2)	COA(1)	COA(2)
核心	18.77	40.88	12.73	27.34	12.73	31.32
一般	4.2	10.82	7.22	16.89	4.49	10.01

反过来，当一个专利的 COA 值确定后，可以对专利进行分类。基于 COA 值，作者设计了一个线性分类器，分类结果如表 10 所示。

表 10　　专利分类结果

方法	数据集 1	数据集 2	数据集 3
Patent Citation	59.75	54.84	51.82
COA(1)	53.20	70.63	64.49
COA(2)	59.76	69.79	60.15

一般来说，一个专利如果被越多的专利引用，则该专利越有价值。通过实验发现，COA 方法比直接利用引用关系评估专利价值准确率高。

3.3　PatentDom 分析框架

PatentDom 是一个基于网络的专利分析框架[37,38]，基于该框架设计了 3 个应用：PatentLine、PatentTrace 和 PatentLink。PatentDom 引入多视图专利图(multi-view patent graph)概念，$G=(V,w_v,E_s,w_s,E_{ct},w_{ct})$。其中，$V$ 对应于每一个专利，每个节点都有一个权值，为该专利被引用次数的倒数，所以权值越小代表该专利越重要。图 G 中包含 2 种类型的边。如果 2 个专利的相似度超过一定的权值，那么它们之间就存在一条无向边，相似度对应于该边的权值。如果专利之间存在引用关系或者 2 个专利发布的时间小于预先设定的时间间隔，那么 2 个节点之间存在一个有向边，每个有向边的权重为 1。由于该网络包含 2 种类型的边，因此称为多视图专利图。

3 个应用的核心是从图 G 中选择 L 个最重要的专利。PatentDom 将此问题归结为图论中最小支配集问题，利用贪心算法，选择起决定性作用的 L 个专利。

PatentLine 主要分析核心专利随时间变化的关系。该框架将问题归结为最小代价的 Steiner 树，利用生成树建立核心专利之间的联系。

PatentTrace 用来分析一个给定的专利 q 和最重要的 L 个专利之间的关系，即分析该专利最大可能和那个重要的专利之间存在关联。PatentTrace 采用式（15）计算节点的权值。

$$cost(v)=\frac{1-sim(v,q)}{citation(v)} \tag{15}$$

该计算方法综合了专利之间文本的相似度和引用关系。

PatentLink 则利用中心子图(center-piece)分析 2 个专利之间潜在的联系。

通过典型案例研究表明，这 3 种分析方法的结果是有效的，能够分析出专利技术发展的脉络。

由于 3 种分析方法都依赖于核心专利的选择，表 11 是 PatentDom 方法、COA[28]方法、PageRank 方法以及 CorePatent 方法检索结果的对比。通过对比可以看出，PatentDom 在目前已有的方法中对专利价值的计算是比较好的。

4　基于深度学习的专利检索与分析

面对海量的专利数据，即使是技术很全面的专利工作者也往往力不从心。由于专利撰写的特点使专利检索的召回率和准确率有待进一步的提高[39]。目前，专利检索与分析主要针对专业人员，一般人很难利用，因此需要专利检索与分析更加准确和智能化，下面本文从专利检索、专利论文检索以及专利趋势分析 3 个方面举例说明深度学习在专利检索与分析中的应用[40,41]。

表 11 核心专利检索结果对比

方法	top@10			top@30			top@50		
	准确率	召回率	*F*1	准确率	召回率	*F*1	准确率	召回率	*F*1
COA	0.11	0.056	0.07	0.092	0.138	0.11	0.086	0.215	0.123
PageRank	0.106	0.053	0.07	0.1	0.15	0.12	0.112	0.28	0.16
CorePatent	0.188	0.094	0.125	0.192	0.288	0.231	0.192	0.48	0.274
PatentDom	0.194	0.097	0.129	0.22	0.33	0.263	0.212	0.53	0.3

4.1 专利检索

图 3 是一个基于深度学习的专利检索方法，对于一个待检索的专利 Q，从专利库中检索类似的专利 $C_1,C_2,\cdots,C_n$。该方法分为 2 步。

1) 特征提取，将专利语料库映射到一个 k 维的空间。对于一个给定的专利，利用卷积神经网络(CovNN，convolutional neural network)将专利文本通过多层卷积，提取其 k 维语义特征[41]。

一篇专利包含标题、摘要、正文（实施）和权利声明等几个部分。如果一个专利包含图表，则还有相应的关于图表的说明。专利每一部分所表达的内容不同，以及申请人在每一部分的撰写方式不同，因此本文认为每一部分存在不同的特征。所以在卷积神经网络的第一层，本文设置了 4 个卷积核 $g_1(x),g_2(x),g_3(x),g_4(x)$，对每一个部分进行初始特征提取。

由于专利每一部分的长度不一样，摘要部分言简意赅，实施部分详细明了，权利声明部分则居于两者之间，因此需要设计每一个卷积函数的步长，每一次卷积操作可以看成是对步长内的文本信息进行特征提取。通过第一层卷积神经网络，本文完成专利文本的原始输入，并提取了初步特征。但这些特征还比较局部，为了进一步提取全局特征并降低输入的维度，需要通过多层卷积神经网络对第一层卷积神经网络的输出进行再次卷积。在每一层卷积网络中，本文同样需要设计多个卷积核，这样可以从不同的角度提取专利文本的特征，当提取多重特征后，需要设计合适的池化(max-pooling)方法对特征进行融合。最终对于给定的专利 P_i，本文得到它的 k 维向量，设为 V_i。

卷积神经网络的参数训练过程可以采用梯度下降的过程进行逐层训练。这里再引入一个相似度

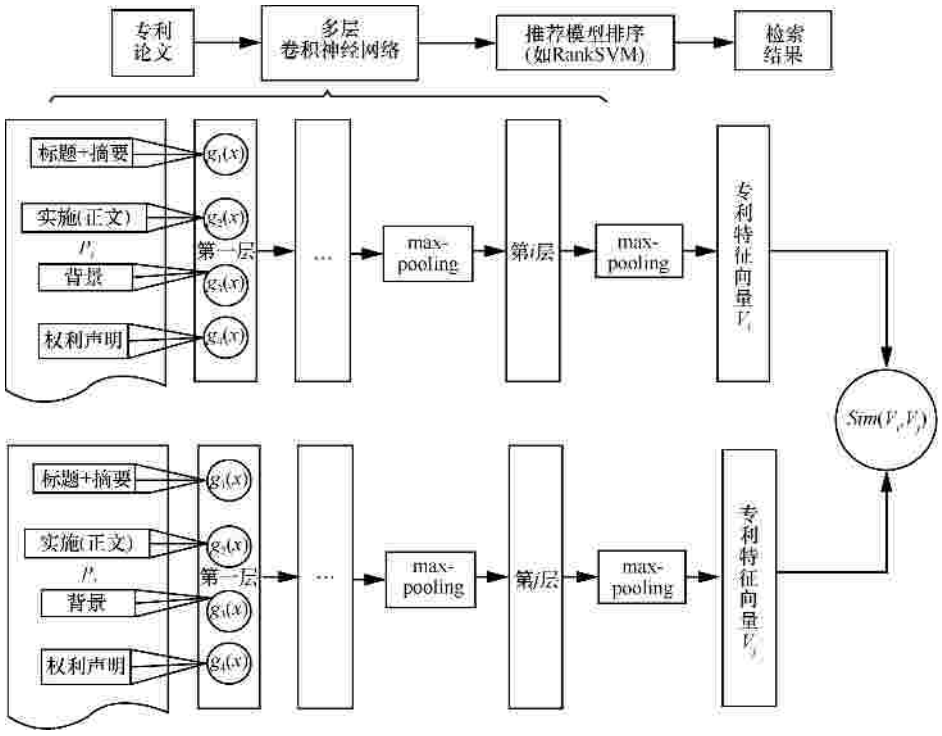

图 3 基于深度学习的专利检索

函数使提取的特征是有效的。由于提取的是专利的语义特征，本文采用传统余弦相似度对 2 个专利进行相似度计算。如果专利 P_i 和 P_j 相似，P_i 和 P_k 不相似，那么 $Sim(V_i, V_j)>>Sim(V_i, V_k)$。如果不等式不成立，那么卷积网络提取的特征是有偏差的，这样本文利用相似度作为目标函数去优化卷积神经网络的卷积核。

2) 利用排序学习的方法，对检索到的专利进行排序。这里排序学习考虑的因素有专利的语义相似度、专利发布的时间、专利的法律状态、专利之间的引用关系等。专利的语义相似度采用余弦相似度进行计算。假设专利 P_i 和 P_j 的发布时间分别是 $Year(P_i)$和 $Year(P_j)$，那么专利之间相对价值采用式(16)计算。其含义是优先推荐最近的专利。

$$Value(P_i, P_j) = e^{Year(P_j)-Year(P_i)} \tag{16}$$

专利之间存在引用关系，这样就可以构造专利引用网络，根据专利在网络中的相对关系，采用网络的度量指标（如距离、跳数）计算专利在技术上的关联程度。这样就可以构造一个排序学习算法向用户推荐最相似的专利。

4.2　专利论文检索

4.1 节主要研究在专利文档集合中检索相似的专利，同样论文也是一个很重要的技术文献集合，论文中包含了大量的技术。

对于一个专利，检索与之相关的论文可以帮助专利审查员决定该专利是否新颖，同样对于一个公司可以帮助公司研发人员掌握更全面的相关领域的技术现状。因此对于一个专利检索相似的论文也是一个值得研究的问题。图 4 是一个基于深度学习的专利论文检索框架。

Step1　特征提取。同样采用卷积神经网络对论文和专利分别提取其相应的特征。由于论文和专利分属不同的科技文献种类，因此需要设计不同的卷积函数对其进行特征提取。

Step2　空间变换。由于论文和专利属于不同类的科技文献，因此可以认为提取的特征属于不同的空间，为了计算其相似程度需要对它进行空间变换。假设 V_i 和 V_j 分别为专利 P_i 和论文 A_j 所对应的 k 维向量(假设为列向量)。本文定义存在一个 $k\times k$ 维的矩阵 $\boldsymbol{M}$，使 $V_i=\boldsymbol{M}V_j$。它的含义是，如果 P_i 和 A_j 是相似的，那么在向量空间存在某种形式的矩阵变换使向量 $\boldsymbol{V}_j$ 变换成 $\boldsymbol{V}_i$。

本文使用目标函数优化的方法计算矩阵 $\boldsymbol{M}$，目标函数如式（17），其中，n 是给定的测试数据集中数据的个数。

$$J = \min \sum_{i=0}^{n} (1 - \cos(\boldsymbol{V}_i, \boldsymbol{V}_j)) \tag{17}$$

目标函数中采用余弦函数，这是因为如果 2 个向量在线性空间越相似，其余弦值越大，$1-\cos(\boldsymbol{V}_i,\boldsymbol{V}_j)$

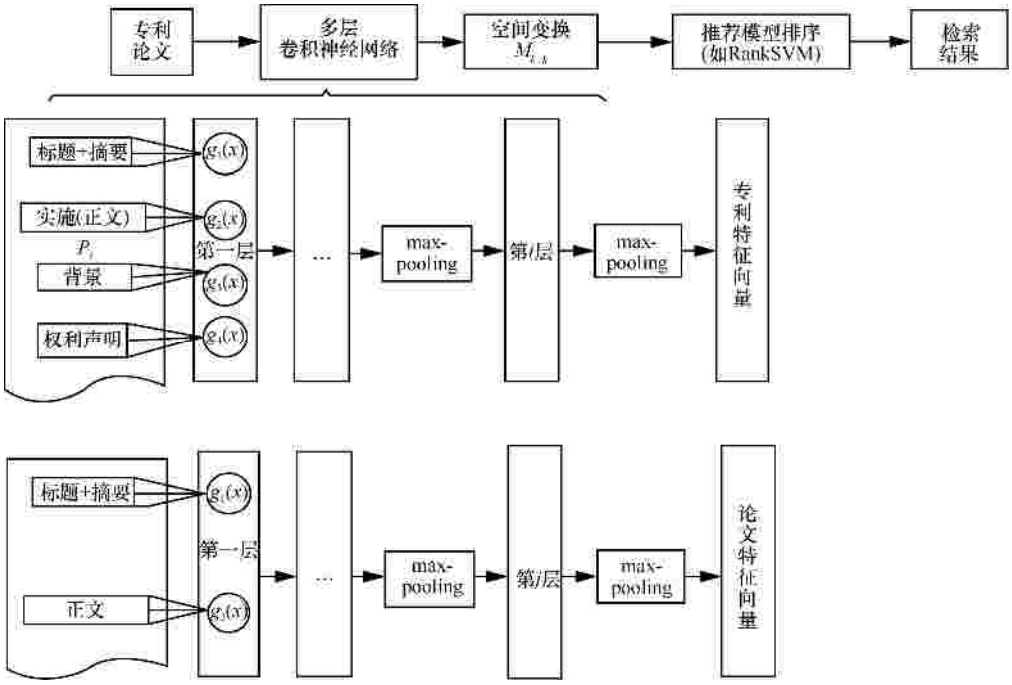

图 4　基于深度学习的专利论文检索框架

越小。对目标函数采用梯度下降的算法对其进行优化，从而得到转换矩阵 M。

Step3 利用排序学习的方法，对检索到的专利进行推荐。这里排序学习考虑的因素有论文的语义相似度、论文发布的时间、论文的质量以及论文之间的引用关系等因素。

4.3 结合论文的专利趋势分析

专利趋势分析就是分析某个领域现有专利技术发展的现状。正如前面提到科技论文也是一个非常重要的技术来源，在分析专利发展趋势时必须考虑科技论文。

在前面已经设计了卷积神经网络提取专利和论文的特征，并构建了专利和论文之间进行特征转换的矩阵 M，这样就可以将论文和专利映射到同一个语义空间。

Step1 利用现有的聚类算法，将专利和科技论文进行聚类。

Step2 对于每一类，利用深度学习提取专利和论文中的技术短语（算法1）。

Step3 对每一类技术短语建立 Logistic 模型，确定其参数，并预测专利的发展趋势。

此外，专利和论文是从不同的方面反映了技术发展的历程。有的领域论文在先，研究人员开展大量的基础研究或者理论研究，到达一定实用阶段时可以去申请大量的专利。有的领域可能是专利在前面，再有大量的研究，如 PageRank 算法。类似产品的生命周期，本文将技术的生命周期分为4个阶段：导入期、成长期、成熟期和衰退期。

在每一类中，分别对论文和专利建立其相应的 Logistic 模型，并分析所处的阶段，建立每一个阶段论文和专利之间的时间对应关系，这样更好地帮助企业去预测技术的发展。

算法1 ExtractTechnicalTerm//提取技术短语

1) 使用公开的语料库建立初始的字向量，向量维度为100，迭代100次。

2) 抽取德温特专利数据库中人工标注的技术短语作为训练数据。

3) 使用左右各4个字做为上下文，9×100个神经元为输入层，隐藏层为100，输出层为4，神经网络结构为[900→100→4]，进行 n 次迭代，建立深度神经网络 DNN-TM[40,41]。

4) 用 DNN-TM 神经网络抽取专利和论文中的技术短语。

这一节研究了专利检索、结合论文的专利检索方法，均采用了卷积神经网络提取专利和论文的特征，避免了文本稀疏带来的“维数灾难”。方法的核心就是确定卷积神经网络的结构：卷积核的个数及其参数、卷积网络的层数。其次，本文设计了空间转换矩阵，利用目标函数优化的方法实现了论文和专利之间的语义转换。

在专利趋势分析中，本文提出利用深度神经网络提取技术短语词汇，利用生命周期模型，建立论文和专利生命周期之间的对应关系，帮助用户更好地预测技术发展的趋势。

5 结束语

国家和企业越来越重视知识产权的保护，研究人员提出专利的技术现状检索和相关性检索等专利检索方法，设计专利新颖度分析和专利地图分析等专利分析方法，使企业用户可以快速地了解领域的研究现状，把握技术趋势变化，做出合理的企业决策。

在专利检索方面，学者们已经取得了丰硕的成果，提出基于主题的检索、基于引用的检索、基于词库的扩展检索等多种检索方法，但是准确率和召回率仍然有待提高。在专利分析方面，尽管已经取得了一定的成果，但对专利数据的分析仍然较浅[39]。如专利中包含的引用关系很少被考虑到，而进行专利搜索与分析的研究时，如果能够结合引用关系，会使检索和分析结果更加准确。此外，专利文献不仅包括中文，还有英文、日文专利等，并且科技论文中同样包含大量的技术，因此本文必须设计新的智能化专利搜索与分析算法，使之能够适应跨语言、跨语料库的专利检索和分析，这样才能够真正发挥它们的巨大作用。

参考文献：

[1] State Intellectual Property Office of PRC. 2014 key IP5 statistical data[EB/OL].http://www.sipo.gov.cn/tjxx/wjndbg/201507/P020150707534432342721.pdf.

[2] State Intellectual Property Office of PRC. 2013 key IP5 statistical data[EB/OL].http://www.sipo.gov.cn/tjxx/wjndbg/201509/P020150901583608432123.pdf.

[3] State Intellectual Property Office of PRC. 2012 key IP5 statistical data[EB/OL]. http://www.sipo.gov.cn/tjxx/2012tjbgen.pdf.

[4] CHEN C. Searching for intellectual turning points: progressive knowledge domain visualization[J]. PNAS, 2004, 1011(Suppll): 5303-5310.

[5] ERDI P, MAKOVI M, SOMOGYVARI Z, et al. Prediction of Emerging technologies based on analysis of the US patent citation network[J].

Scientometrics, 2013, 95(1): 225-242.
[6] FUJII A, ISHIKAWA T. NTCIR-3 patent retrieval experiments at ULIS[C]//NII Test Collection for IR Systems-3. c2002: 1-6.
[7] FUJII A, ISHIKAWA T, KANDO N. Test collections for ptent-to-ptent rtrieval and ptent map generation in NTCIR-4 workshop[C]//The 4th International Conference on Language Resources and Eva uation. c2004: 1643-1646.
[8] FUJII A, ISHIKAWA T, KANDO N. Overview of the patent retrieval task at the NTCIR-6 workshop[C]//NII Test Collection for IR Systems-6. Tokyo, Japan, c2007: 359-365.
[9] WU S, SUN J, TANG J. Patent partner recommendation in enterprise social networks[C]//WSDM, Rome, Italy, c2013: 43-52.
[10] JIN X, SPANGLER S, CHEN Y, et al. Patent maintenance recommendation with patent information network model[C]//ICDM. Vancouver, Canada, c2011: 280-289.
[11] MANNING C, RAGAVAN P, SCHUTZE H. An introduction to information retrieval[M]. London: Cambridge University Press, 2009.
[12] MAGDY W, JONES G. PRES: a score metric for evaluating recall-oriented information retrieval applications[C]//SIGIR. Geneva, Switzerland, c2010: 611-618.
[13] HIDEO I, HIROKO M, YASUSHI O. Term distillation in patent retrieval[C]//The ACL-2003 Workshop on Patent Corpus. c2003: 41-45.
[14] VERBERNE S，HONDT E D. Prior art retrieval using the claims section as a bag of words[C]//The Cross-language Evaluation Forum Conference on Multilingual Information Access Evaluation: Text Retrieval Experiments. c2009: 497-501.
[15] VARMA M, VARMA V. Applying key phrase extraction to aid invalidity search[C]//International Conference on Artificial Intelligence and Law. Pittsburgh, PA, c2011: 249-255.
[16] KONISHI K. Query terms extraction from patent document for invalidity search[C]//NTCIR-5 Workshop Meeting. Tokyo, Japan, c2005.
[17] MAHDABI P, ANDERSSON L, Keikha M, et al. Automatic refinement of patent queries using concept importance predictors[C]// SIGIR. Portland, USA, c2012: 505-514.
[18] TAKAKI T, FUJII A, ISHIKAWA T. Associative document retrieval by query subtopic analysis and its application to inva idity patent search[C]//CIKM. Washington, USA, c2004: 399-405.
[19] ADAMS S. Comparing the IPC and the US classification systems for the patent searcher[J]. World Patent Information, 2001, 23(1): 15-23.
[20] MAHDABI P, GERANI S, HUANG J X, et al. Leveraging conceptual lexicon: query disambiguation using proximity informat n for patent retrieval[C]//SIGIR. Dublin, Ireland, c2013: 113-122.
[21] GANGULY D, LEVELING L, MAGDY W, et al. Patent query reduction based on pseudo-relevant documents[C]//CIKM. Glasgow, Scotland, UK, c2011: 1953-1956.
[22] MAGDY W, JONES G. A study on query expansion methods for patent retrieval[C]//PAIR. c2011: 19-24.
[23] KRESTEL R, SMYTH P. Recommending patents based on latent topics[C]//Recommender Systems. c2013: 395-398.
[24] MAHDABI P, CRESTANI F. Query-driven mining of citation networks for patent citation retrieval and recommendation[C]//CIKM. Shanghai, China, c2014: 1659-1668.
[25] HIRONORI D, YOHEI S, et al. A patent retrieval method using a hierarchy of clusters at TUT[C]//NTCIR-5 Workshop Meeting. Tokyo, Japan, c2005.
[26] BHATIA S, HE B, HE Q, et al. A scalable approach for performing proximal search for verbose patent search queries[C]//CIKM. Maui, HI, USA, c2012: 2603-2606.
[27] BASHIR S, AUBER A. Analyzing document retrievability in patent retrieval settings[C]//DEXA. c2009: 753-760.
[28] BASHIR S, AUBER A. Improving retrievability of patents in prior-art search[C]//ECIR. Dublin, Ireland, c2010: 457-450.
[29] SOOYOUNG O, ZHEN L, LEE W C, et al. CV-PCR: a context-guided value-driven framework for patent citation recommendation[C]// CIKM. San Francisco, CA, USA, c2013: 2291-2296.
[30] HUANG W, KATARIA S, CARAGEA C, et al. Recommending citations: translating papers into references[C]//CIKM. Maui, HI, USA, c2012: 1910-1914.
[31] XUE X，CROFT W. Automatic query generation for patent search[C]//CIKM. Hong Kong, China, c2009: 2037-2040.
[32] JUN S H, PARK S, SIK J D. Technology forecasting using matrix map and patent clustering[J]//Industrial Management & Data Systems. 2012, 112(5): 786-806.
[33] CHEN X, PENG Z, ZENG C. A co-training based method for chinese patent semantic annotation[C]//CIKM. Maui, HI, USA, c2012: 2379-2382.
[34] LIU D, PENG Z, LIU B. Technology effect phrase extraction in Chinese patent abstracts[C]//APWeb. Changsha, China, c2014: 141-152.
[35] DRAZIC M, KUKOLJ D, VITAS M, et al. Technology matchin of the patent documents using clustering algorithms[C]//The 14th IEEE International Symposium on Computational Intelligence and Informatics. c2013: 405-408.
[36] HASAN M A, SPANGLER S, GRIFFIN T, et al. COA: finding vel patents through text analysis[C]//SIGKDD. Paris, France, c2009: 1175-1184.
[37] ZHANG L H, LI L, LI T, et al. PatentLine: analyzing technology evolution on multi-view patent graphs[C]//SIGIR. Boston, Massachusetts, USA, c2014: 1095-1098.
[38] ZHANG L H, LI L, LI T, et al. PatentDom: analyzing patent relationships on multi-view patent graphs[C]//CIKM. Shanghai, China, c2014: 1369-1378.
[39] TADURI S，YU H，LAU G, et al. Developing a comprehensive patent related information retrieval tool[J]. Journal of Theoretical and Applied Electronic Commerce Research. 2001, 6(2): 1-16.
[40] BNEGIO, Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research. c2009: 1137-1155.
[41] WANG M X, LU Z D, LI H, et al. GenCNN: a convolutional architecture for word sequence prediction[C]//ACL. c2015.

作者简介：

刘斌（1975-），男，江苏泰兴人，博士，武汉大学讲师，主要研究方向为复杂数据管理、数据挖掘等。

冯岭（1986-），男，河南郑州人，武汉大学博士生，主要研究方向为专利分析与挖掘等。

王飞（1989-），男，江苏连云港人，武汉大学博士生，主要研究方向为专利检索、分析和挖掘。

彭智勇（1963-），男，湖北武汉人，武汉大学教授、博士生导师，主要研究方向为复杂数据、可信数据和 Web 数据管理。